Do essential genes evolve slowly?

Laurence D. Hurst and Nick G.C. Smith

Approximately two thirds of all knockouts of individual mouse genes give rise to viable fertile mice. These genes have thus been termed 'non-essential' in contrast to 'essential' genes, the knockouts of which result in death or infertility. Although non-essential genes are likely to be under selection that favours sequence conservation [1], it is predicted that they are less subject to such stabilising selection than essential genes, and hence evolve faster [2]. We have addressed this issue by analysing the molecular evolution of 108 non-essential and 67 essential genes that have been sequenced in both mouse and rat. On preliminary analysis, the non-essential genes appeared to be faster evolving than the essential ones. We found, however, that the non-essential class contains a disproportionate number of immune-system genes that may be under directional selection (that is, selection favouring change) because of host–parasite coevolution. After correction for this bias, we found that the rate at which genes evolve does not correlate with the severity of the knockout phenotype. This was corroborated by the finding that, whereas neuron-specific genes have significantly lower rates of change than other genes, essential and non-essential neuronal genes have comparable rates of evolution. Our findings most probably reflect strong selection acting against even very subtle deleterious phenotypes, and indicate that the putative involvement of directional selection in host–parasite coevolution and gene expression within the nervous system explains much more of the variance in rates of gene evolution than does the knockout phenotype.

Address: Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK.
E-mail: l.d.hurst@bath.ac.uk

Received: **9 April 1999**
Revised: **10 May 1999**
Accepted: **4 June 1999**

Published: **5 July 1999**

Current Biology 1999, 9:747–750
http://biomednet.com/elecref/0960982200900747

Results and discussion

Wilson *et al.* [2] proposed in 1977 that two proteins subject to the same level of functional constraint, but differing in their dispensability, will evolve at different rates. They argued that this may, in part, result from the fact that mutations in redundant genes may be masked. Thus, essential genes (knockouts of which are lethal or infertile) should evolve slower than non-essential genes (knockouts of which are viable and fertile) — this is the 'knockout-rate' prediction. From the adaptive theory of mutation rates, it might similarly be predicted that the mutation rates of essential genes should be lower than those of non-essential genes [3–5]. We have examined these issues using sequence comparison of mouse and rat genes with cross-reference to knockout data (see Materials and methods for details). For any aligned pair of orthologous genes we can calculate K_A, the rate of DNA substitutions affecting the amino-acid composition of the gene product (non-synonymous substitutions), and K_S, the rate of DNA substitutions that are silent at the amino-acid level (synonymous substitutions). If synonymous mutations in mammals are neutral, as may be the case [4], then K_S is a measure of the mutation rate [6] and K_A/K_S is a measure of the rate of protein evolution after controlling for mutation rate.

Rates of evolution and the knockout phenotype

Essential genes ($n = 67$ in our sample) had significantly lower K_A/K_S ratios than non-essential genes ($n = 108$) (Mann-Whitney U-test, $p = 0.0009$, Figure 1; all results and statistics are given in Table 1). Essential genes also tended to have lower K_S values than non-essential genes ($p = 0.04$ in a one-tailed test).

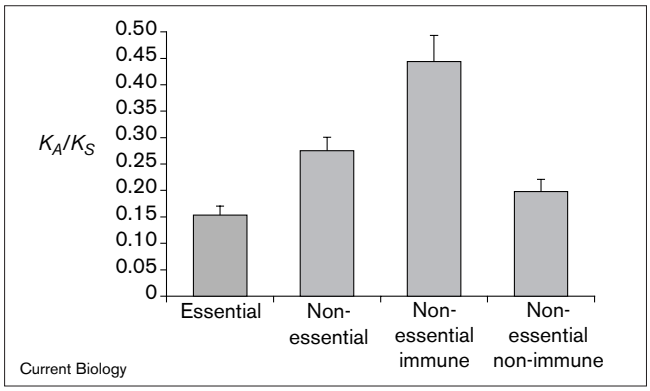
Rates of evolution after controlling for 'tissue of activity'

The above results are potentially confounded by differences between essential and non-essential genes in the types of tissues that they act in. Notably, the non-essential class contained many more genes specific to the immune system than did the essential class (34/108 versus 3/67, $p < 10^{-5}$ using Fisher's exact test). Genes of the immune system tend to have high K_A/K_S ratios [7,8] and high K_S values [7]. We confirmed these results: K_A/K_S ratios of immune-system genes ($n = 37$) were on average more than double those of non-immune genes ($n = 138$, $p = 0.0028$) and immune-system genes also had higher K_S values ($p = 0.02$) (Table 1). The high K_A/K_S ratios can be accounted for by arguing that immune-system genes are, at least along part of their sequence, likely to be under directional selection (and/or overdominance) driven by host–parasite coevolution (see for example [9–11]).

We then asked whether the non-essential non-immune genes ($n = 74$) had higher K_A/K_S and higher K_S values than essential non-immune genes ($n = 64$). When we did this comparison we found no significant differences ($p = 0.24$ and $p = 0.45$, respectively; Table 1).

The effect of tissue specificity on rates of gene evolution can also be controlled for by examining genes expressed specifically in neurons. Neither the K_A/K_S ratios nor the

Figure 1



The K_A/K_S ratios (± SEM) for various classes of gene. The left-hand two columns give the values for essential genes ($n = 67$) and non-essential genes ($n = 108$), respectively. These are significantly different ($p < 0.001$). The right-hand two columns show the data for the non-essential genes broken down into those with immune function ($n = 34$) and those without ($n = 74$). Note that the major difference between the essential and non-essential genes is largely due to the excess of immune-related genes in the latter class.

K_S values for non-essential neuronal genes ($n = 18$) were significantly different from those of the essential neuronal genes ($n = 16$; $p = 0.63$ and $p = 0.36$, respectively). The comparison between all neuronal genes ($n = 34$) and non-immune non-neuronal genes ($n = 104$) revealed that K_A/K_S values for neuronal genes were less than half those of the others ($p = 0.0001$). The two classes had similar K_S values ($p = 0.51$).

Is this low rate of evolution a peculiarity of neuronal genes or a general property of tissue-specific genes? This is hard to evaluate, but the fact that genes specifically involved in reproduction ($n = 18$) had higher K_A/K_S ratios than neuronal genes ($n = 34$, $p = 0.013$) suggests that the low rate is peculiar to neuronal genes. This was further corroborated by the finding that reproductive genes had similar K_A/K_S ratios and K_S values to those of the 'normal' class of non-immune non-neuronal genes ($p = 0.91$ and $p = 0.31$, respectively).

The effect of tandem substitutions

In the mouse–rat comparison there is a strong correlation between K_A and K_S (see for example [12,13]). It seems that this correlation is in large part due to an excess of tandem (adjacent) substitutions, which may be the result of selection (N.G.C.S. and L.D.H., unpublished observations). If selection acting on synonymous mutations is responsible for the excess of tandem substitutions, then K_S is not an unbiased measure of the mutation rate. We therefore re-analysed our sequences ignoring tandem substitutions.

Our results concerning stabilizing selection were not affected when tandem substitutions were ignored. The non-essential genes had a significantly higher K_A/K_S ratio than the essential genes ($p = 0.0007$). This result was again attributable to the excess of immune-system genes within the non-essential class, as the K_A/K_S ratios are similar for non-immune essential and non-immune non-essential genes ($p = 0.288$). Immune-system genes had high K_A/K_S ratios compared with all non-immune genes ($p < 0.0001$). Neuronal genes had especially low K_A/K_S ratios compared with non-immune non-neuronal genes ($p < 0.0001$). In contrast, our results as regards K_S were altered when tandem substitutions were ignored. The essential genes had mutation rates no lower than those of non-essential genes ($p = 0.25$) and the immune genes had K_S values no higher than non-immune genes ($p = 0.44$). Thus, the variation in 'raw' K_S values might partly have been due to selection on synonymous mutations. With the effects of such selection removed, we found no evidence of adaptive variation in mutation rates. Hence, we need only seek to explain the variation in K_A/K_S ratios.

Why is there no correlation between knockout phenotype and rates of evolution?

Our results suggest that knockout phenotype does not correlate with the K_A/K_S ratio once we have accounted for tissue-specific effects. To explain this result it is worth clarifying the logic underlying the knockout-rate prediction. The knockout phenotype indicates the magnitude of the selective difference between the presence and absence of the gene: the more severe the knockout phenotype, the greater the strength of selection on the entire gene. The knockout-rate prediction requires not only that the strength of selection on the entire gene should co-vary with the average strength of selection on non-synonymous mutations within the gene (although each mutation is likely to be under weaker selection than the null mutation) but, crucially, that the strength of selection (s) on these individual mutations (the magnitude of their effects on fitness) is within certain bounds.

Imagine that non-synonymous mutations in non-essential genes tend to have, on average, a smaller effect on fitness than do mutations in essential genes. This will not lead to differences in the rate of fixation of mutations if the mutations are highly deleterious: in a reasonably sized population a mutation reducing fitness by 10% will be no more likely to reach fixation than one that reduces fitness by 20%. But for very low s in a finite population, however, deleterious mutations can reach fixation (become a substitution) by random drift. In this case, the size of s affects the fixation probability and hence the substitution rate. So one way of arriving at the knockout-rate prediction is to suppose that non-essential genes have more non-synonymous mutations with s values that are compatible with fixation. Our failure to find any differences in rates of evolution between essential and non-essential genes might then be due to the fact that most non-synonymous mutations in the two types of genes are highly deleterious

Table 1

The *p* values from two-tailed Mann-Whitney U-tests in pairwise comparisons of various gene classes.

Class	*n*	K_A/K_S		NE	E	I-NE	I	NI-E	NI-NE	NR	NNR-NI	NR-E	NR-NE	NI	R
			K_S	0.186 ± 0.007	0.165 ± 0.007	0.209 ± 0.014	0.204 ± 0.013	0.165 ± 0.008	0.177 ± 0.008	0.163 ± 0.011	0.174 ± 0.006	0.148 ± 0.014	0.176 ± 0.017	0.171 ± 0.005	0.152 ± 0.016
Non-essential (NE)	108	0.275 ± 0.024		–	0.0763	*0.1495*	*0.2411*	0.0904	*0.3878*	*0.1384*	*0.2410*	0.0838	*0.5890*	*0.1321*	*0.1143*
Essential (E)	67	0.153 ± 0.017		**0.0009**	–	**0.0062**	***0.0114***	*0.9725*	0.4172	*0.9227*	*0.4854*	*0.4706*	0.6019	*0.6042*	*0.5944*
Immune non-essential (I-NE)	34	0.444 ± 0.048		***0.0005***	**0.0000**	–	*0.7779*	**0.0080**	**0.0461**	**0.0170**	**0.0226**	**0.0115**	0.1632	**0.0116**	**0.0205**
Immune (I)	37	0.418 ± 0.047		***0.0023***	***0.0000***	*0.6166*	–	**0.0143**	0.0762	**0.0275**	**0.0399**	**0.0176**	0.2192	**0.0212**	**0.0307**
Non-immune essential (NI-E)	64	0.155 ± 0.018		**0.0012**	*0.9962*	**0.0000**	**0.0000**	–	0.4460	*0.9020*	*0.5189*	*0.4595*	0.6181	*0.6405*	*0.5792*
Non-immune non-essential (NI-NE)	74	0.198 ± 0.022		***0.0356***	0.2301	**0.0000**	**0.0000**	0.2445	–	*0.4468*	*0.8560*	0.2457	*0.9647*	*0.6776*	*0.2991*
Neuronal (NR)	34	0.096 ± 0.02		***0.0000***	***0.0196***	**0.0000**	**0.0000**	***0.0220***	*0.0014*	–	0.5127	*0.5815*	*0.6103*	*0.6104*	0.7583
Non-neuronal non-immune (NNR-NI)	104	0.20 ± 0.017		*0.0661*	*0.0610*	**0.0000**	**0.0000**	*0.0694*	*0.5963*	**0.0001**	–	0.2613	0.9281	*0.8059*	*0.3104*
Neuronal essential (NR-E)	16	0.081 ± 0.019		**0.0005**	*0.0934*	**0.0000**	**0.0000**	*0.1006*	**0.0145**	0.7640	**0.0037**	–	0.3605	0.3070	0.8360
Neuronal non-essential (NR-NE)	18	0.109 ± 0.038		***0.0005***	0.0550	**0.0000**	**0.0000**	*0.0578*	***0.0141***	*0.7902*	**0.0040**	0.6255	–	*0.8266*	0.4765
Non-immune (NI)	138	0.177 ± 0.015		***0.0013***	*0.4894*	**0.0000**	**0.0028**	*0.5066*	*0.4887*	***0.0027***	*0.1552*	***0.0313***	***0.0211***	–	*0.3838*
Reproductive (R)	18	0.193 ± 0.037		*0.2676*	*0.2781*	**0.0009**	**0.0028**	*0.2882*	*0.8982*	**0.0130**	*0.9138*	**0.0329**	**0.0354**	*0.5371*	–

In the lower left half of the table are the *p* values for each of the pairwise K_A/K_S comparisons between all of the gene classes and in the upper right half are the *p* values for the K_S comparisons. The sample size (*n*) and the mean K_A/K_S and K_S values (± SEM) for each gene class are given. Significant *p* values are shown in bold. Those shown in italics are comparisons in which there is overlap between data sets. The genes in this analysis have not had tandem substitutions removed. A significant statistic indicates that the set of genes in one class is significantly greater or smaller in the parameter in question than the other set of genes. To establish which is the greater, it is necessary to consult the absolute mean values for K_A/K_S or K_S given in column 3 and row 2, respectively.

(that is, selection is very strong), and that the proportion of slightly deleterious or effectively neutral mutations in the two is about the same.

There may, however, be other explanations. First, there may be methodological problems. The knockout phenotype may be a poor indicator of the fitness effect of a knockout in nature, owing to the artificiality of the laboratory environment. Alternatively, the mouse–rat data may be misleading. Mutational saturation of sites is a potential problem in any analysis of this sort. A sizeable majority of silent sites in the mouse–rat comparison have not changed, however, indicating that saturation is unlikely. Second, some of the knockout-rate prediction's other assumptions may be invalid. The prediction also requires, for example, that advantageous mutations are rare. Analysis of non-synonymous evolution in immune-system genes and in hemizygously expressed genes [4,5] indicates that this may be unrealistic. How this might affect the predictions is unclear.

Why is there a correlation between tissue specificity and rates of evolution?

We have found that the rate of evolution as measured by the K_A/K_S ratio is strongly affected by tissue specificity: immune-system genes evolve rapidly, neuronal genes evolve slowly and reproductive genes evolve at an intermediate rate. There are a number of possible explanations for an effect of tissue specificity on K_A/K_S values. Genes in different tissues might by affected by different proportions

of advantageous and deleterious mutations. A high proportion of advantageous mutations would seem to explain the high rates of evolution of immune-system genes. Tissues may also differ in their influence on an organism's fitness, although this is, as yet, mere speculation. But we can be sure that the putative involvement in host–parasite coevolution and specific expression within the nervous system explain the variance in the rate of protein evolution much better than do knockout phenotypes.

Materials and methods

Database assembly

The gene knockout database (gkd: www.bioscience.org/knockout/knochome.htm) annotates the phenotypes of over 300 single-gene knockouts in mice. We also used four genes annotated in Tbase (tbase.jax.org/) but not present in gkd. Knockouts reported as infertile or inviable were incorporated into the class of essential genes. Genes whose knockout had sex-specific infertility or high, but not complete, levels of inviability were also included within this class. Where some phenotype was observed (in 97% of cases) the genes are classified in the gkd by the tissue principally disrupted. Some are classed as multi-tissue if no one organ or system is predominantly affected. We checked all genes against the relevant Tbase entries or original papers to confirm the designation of the 'tissue of activity' and the viability/fertility. Internet links to the genome database (gdbwww.gdb.org/gdb/), mouse genome informatics (mgd.hgmp.mrc.ac.uk/), Online Mendelian Inheritance in Man (www3.ncbi.nlm.nih.gov/omim/) or to the original paper, were followed to obtain links to the GenBank citation for the gene concerned.

For each gene the HOVERGEN database of vertebrate homologous genes [14] was searched using the GenBank entry name. From observation of the phylogeny of each gene family, we assembled a list of accession numbers for the mouse and rat orthologues corresponding to known gene knockouts. The accession numbers of the 108 non-essential and 69 essential genes are provided in the Supplementary material.

Preparation of alignments

FETCH was used to extract sequences from databases and GENETRANS was used to extract and combine exons automatically. Protein alignments were performed using CLUSTALW [15]. The DNA alignments were re-created from the protein alignments and the original DNA sequences using the program MRTRANS (written by W. Pearson, and available at HGMP (www.hgmp.mrc.ac.uk/)). Alignments were performed using the GCG [16] and EGCG [17] packages at HGMP. On close inspection two of the essential genes were found to have dubious alignments and were eliminated, reducing the sample size to 67. For these and the 108 non-essential genes we calculated the number of gaps in the alignments, in order to ask whether the alignment protocol might be affecting our rate estimates. Using both the total number of gaps and the number of gaps per informative site, we found not even a weak correspondence with any of the rate parameters. Alignment artefacts are probably not producing systematically biased rate estimates.

Algorithmic rate estimation

Substitution rates were estimated from alignments using methods developed by Moriyama and Powell [18]. Tamura and Nei's [19] multiple hits correction method was used to account for variation in base composition, and Li's [20] method was used to calculate the non-synonymous rate per site (K_A) and the synonymous rate per site (K_S).

Supplementary material

Supplementary material including a complete list of the genes analysed is available at http://current-biology.com/supmat/supmatin.htm.

Acknowledgements

We wish to thank Mike Danson and Adam Eyre-Walker for discussion.

References

1. Brookfield JFY: **Genetic redundancy: screening for selection in yeast.** *Curr Biol* 1997, **7:**R366-R368.
2. Wilson AC, Carlson SS, White TJ: **Biochemical evolution.** *Annu Rev Biochem* 1977, **46:**573-639.
3. Cox EC: **On the organization of higher chromosomes.** *Nat New Biol* 1972, **92:**133-134.
4. McVean GT, Hurst LD: **Evidence for a selectively favourable reduction in the mutation rate of the X chromosome.** *Nature* 1997, **386:**388-392.
5. Smith GC, Hurst LD: **The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate?** *Genetics* 1999, in press.
6. Kimura M: *The Neutral Theory of Evolution.* Cambridge, UK: Cambridge University Press; 1983.
7. McVean GT, Hurst LD: **Molecular evolution of imprinted genes: no evidence for antagonistic coevolution.** *Proc R Soc Lond B* 1997, **264:**739-746.
8. Kuma K, Iwabe N, Miyata T: **Functional constraints against variations on molecules from the tissue level — slowly evolving brain-specific genes demonstrated by protein-kinase and immunoglobulin supergene families.** *Mol Biol Evol* 1995, **12:**123-130.
9. Hughes AL: **Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells.** *Mol Biol Evol* 1997, **14:**1-5.
10. Hughes AL, Nei M: **Pattern of nucleotide substitution at major histocompatibility complex class I loci: evidence for overdominant selection.** *Nature* 1988, **335:**167-170.
11. Hughes AL, Ota T, Nei M: **Positive Darwinian selection promotes charge profile diversity in the antigen binding cleft of class I MHC molecules.** *Mol Biol Evol* 1990, **7:**515-524.
12. Wolfe KH, Sharp PM: **Mammalian gene evolution: nucleotide sequence divergence between mouse and rat.** *J Mol Evol* 1993, **37:**441-456.
13. Makalowski W, Boguski MS: **Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes.** *J Mol Evol* 1998, **47:**119-121.
14. Duret L, Mouchiroud D, Gouy M: **Hovergen — a database of homologous vertebrate genes.** *Nucl Acid Res* 1994, **22:**2360-2365.
15. Thompson JD, Higgins DG, Gibson TJ: **Clustal-W — improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acid Res* 1994, **22:**4673-4680.
16. GCG: *Program Manual for the Wisconsin Package,Version 8.* Wisconsin, USA: Genetics Computer Group; 1994.
17. Rice P: *Program Manual for the EGCG Package.* Cambridge, UK: The Sanger Centre; 1997.
18. Moriyama EN, Powell JR: **Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes.** *J Mol Evol* 1997, **45:**378-391.
19. Tamura K, Nei M: **Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.** *Mol Biol Evol* 1993, **10:**512-526.
20. Li W-H: **Unbiased estimation of the rates of synonymous and nonsynonymous substitution.** *J Mol Evol* 1993, **36:**96-99.